Random verses Non-Random Assignment:
Reply to The GAO Report on
BOP Education and Work Training Programs

GAO's critique of the Post Release Employment Project (PREP) refers to several factors which limit the conclusiveness of the study. This reply addresses the two most significant factors, PREP's non-experimental design and the extent to which the study findings can be generalized. While the GAO report credits us with having acknowledged these limitations, it is misleading because it failed to accurately convey our statements explaining why we designed the study as we did.

As we point out in our PREP reports, when possible a randomized experimental design is preferable to a non-randomized design. However, a randomized design is not always feasible. In many instances the scientist cannot control research group assignment (that is, assignment to either a study or control/comparison group). PREP is one such instance. The limits to the conclusiveness of the findings are not, as GAO suggested, due to the absence of a randomized experimental design. In fact, a randomized experimental design simply is inappropriate for studies like PREP. An experimental design would not mitigate any of the limitations of the study and could even compound the problem by allowing the researchers to believe they have controlled for something that is, in reality, out of their control. This false sense of control could reduce the pursuit of alternative means of controlling for a lack of comparable research groups and lead to a more biased estimate of the treatment effect.

Beyond the ethical and practical problems that would arise from randomly assigning inmates to programs for evaluation purposes, it is naive to believe that individuals would become meaningfully engaged participants simply because they had been assigned to a program. In other words, individuals ultimately control whether they wish to become involved in a program by self-selecting themselves into or out of the program despite any random assignment that might precede the individual's decision.

Most of the existing experimental designs have their origin in agricultural applications, to ascertain the genetics of seeds or the effectiveness of farming methods or fertilizers on crop yields. In this context, when a corn seed is randomly assigned to a parcel of dirt it will grow or not depending on its genetic makeup, while any confounding influences are mitigated via the random assignment to a particular area of a field. These experimental designs have been applied to other substantive areas, many of which involve human subjects. For example, in pharmaceutical studies wherein some individuals are given and ingest a particular drug and the

scientist observes whether the drug has an effect.

Applying these randomized designs to human subjects in the context of social program evaluations, however, can pose a different set of problems than those encountered in applying experimental designs to human subjects in, for example, the medical field. All applications of experimental designs to human subjects share the necessity of controlling for myriad physical and social differences in human subjects, because these individual differences can result in a self-selection process that ultimately determines an individual's research group assignment. But, the self-selection problem is less tangible in applications to social program evaluations. Unlike agricultural applications wherein, for example, all the study seeds are in reality treated in some way, or in a medical application wherein a subject actually receives a treatment, whether or not all of the subjects randomly assigned to a study group are exposed to the treatment is not as easily determined.

In employing an experimental design to a program evaluation, the scientist is frequently left with two choices, treat every individual assigned to the study group as though they received treatment, even though some may not have; or treat only those randomly assigned to the study group who, based on some criterion, are deemed to have actually received the program as study observations and exclude those who self-selected themselves out of the treatment. The scientists dilemma is that when the experimental design fails, there is no good solution as to how to proceed. That is, presuming that every individual identified as a study group member received the treatment is no less problematic than excluding those who did not receive the treatment because they self-selected themselves out of the treatment. In either instance our understanding of the program effect is biased.

The result of a randomized experimental design that does not acknowledge individuals who drop out of their assignment to the study group, is that nonparticipating members of the study group dilute the estimation of the treatment effect.

Alternatively, if the group is composed of a subset of the group which would have existed had the assignment occurred from a totally self-selected set of individuals. Assuming that the control or comparison group membership required no participation or cooperation it would be composed of a random sample of inmates from the general population. The experimental design could, therefore, produce research groups that are not directly comparable because they differ systematically with respect to characteristics that are related to the outcome measure.

There are, therefore, two significant impediments related to the

use of experimental designs in the evaluation of social programs. First, how one handles the problem of self-selection bias. Second, social programs can only have an impact on individuals who are willing to participate in the program. Therefore, an adequate research design must generate a comparison group that is composed of individuals who have a strong likelihood of engaging in the treatment if it is made available to them.

Given this reality, there are two means of controlling for the systematic differences between research groups which arise in situations where randomized experimental designs are not feasible: 1) compare individuals who have been matched based on observed characteristics that are believed to influence the outcome measure, and 2) statistically adjust for the differences in the groups before making any comparisons. In PREP, both methods were employed.

The matching was accomplished by mathematically modeling the self-selection process to produce a group that looks as much like the study group as possible. The PREP research demonstrated that the study and comparison groups were virtually indistinguishable from one another with respect to a wide array of measures (prior incarceration educational, occupational and criminal histories) that one would expect to influence not only their decision to participate in UNICOR, but more importantly their likelihood for success after release. (Both groups differ considerably from the general inmate population.) Arguably, if the individuals in the research groups were indistinguishable when they arrived at the Bureau, then it seems plausible that at least part of the difference observed in the outcomes of the group members could be attributable to their experiences while incarcerated.

With respect to the design of the control group for the PREP, we feel confident that there were many individuals in the population who had an interest in working in prison industries, and would have, had the opportunity been available. Throughout the duration of the PREP about 35% of the inmates housed in Bureau facilities were employed by prison industries, however the waiting list to become employed by prison industries was always substantial. There were always far more inmates who desired a prison industries job than prison industries could accommodate

Finally, with regard to the inability to generalize the findings to the entire inmate population, there is no question, the results do not generalize. As noted above, the individuals who self-selected themselves into the study group were very different from the general inmate population. This is precisely the reason that the matching method was utilized, to define a subset of the general population that looked like the study population with respect to an array of relevant and available measures. In other words, an

experimental design would have produced two groups of individuals that would greatly overestimate the program effect; the method we used, however, because it acknowledged the differences between study members and the general inmate population, resulted in a far more conservative and realistic estimate of the program effect.

Another issue concerning the extent to which the results of the PREP could be generalized to the inmate population pertains to the fact that our study group was composed primarily of inmates employed by UNICOR. Even if the results cannot generalized beyond this 30 to 35 percent of the inmate population, the implications of the study relate to a sizable number of individuals in both absolute and relative terms.

William G. Saylor
Office of Research and Evaluation
(202) 724-3121

March 12, 1993
Revised December 16, 1996